­

**Exploratory Data Analysis: Churn Data**

Stephen E. Porter

College of Information Technology, Western Governors University

Dr. William Sewell

November 11, 2021

**Exploratory Data Analysis: Churn Data**

In this paper, I will use a data set containing cleaned customer data from a fictional telecommunications company. The primary purpose is to perform statistical analysis on the cleaned data set to determine which factor (or factors) are the greatest indicator of customer turnover.

# Part I: Research Question

What factor (or factors) are common among customers who have cancelled their service? By identifying patterns in customers who have terminated their service, organization stakeholders will be able to make changes in their company to 1) correct any flaws within the service model that may lead a customer to leave, and/or 2) identify trends in customers who have left and use that to target customers with similar profiles who may be at risk of leaving. If left unchecked, negative experiences can cause issues beyond losing current customers to negative word-of-mouth stories that drive away potential new customers. (AltlexSoft, 2020)

Relevant data to answer this question come from the Principal Component Analysis (PCA) I performed in the previous course, D206 Data Cleaning. I grouped the columns into each Principal Component (PC) as follows:

* PC1: **Customer Service** (Response, Fix, Replacement, Respectful, Courteous, Listening)
* PC2: **Location 1** (Zip, Lng)
* PC3: **Tenure/Bandwidth** (Tenure, Bandwidth\_GB\_Year)
* PC4: **Services** (Reliability, Options)
* PC5: **Location 2** (Lat, Population)
* PC6: **Family/Support** (Children, Age, Contacts)
* PC7: **Emails/Equipment** (Email, Yearly\_equip\_failure)
* PC8: **Finances** (Income, MonthlyCharge)

These values will be compared against the Churn Data (Yes/No) column in the data frame.

# Part II: Data Analysis

## Statistical Tests

I chose to perform the chi-squared test of independence for the categorical variables found in PC1: Customer Service, PC4: Services, PC6: Family/Support, and PC7: Emails/Equipment. Chi-squared is inappropriate for analysis of the continuous variables from PC2: Location 1, PC3: Tenure/Bandwidth, PC5: Location 2, and PC8: Finances so those will not be analyzed.

I will provide an overview of my process and findings here. Please see the included *R* file: *churn\_data\_analysis.R* for the complete code.

## Principal Component 1

For PC1, I first visualized each variable based on the Yes/No value of Churn. Based on the visualizations in Figure 1 (above), there did not appear to be a strong indication that any of the variables in PC1 influence Churn. I then performed chi-squared tests for each variable to confirm this conclusion using p-values. The p-values found were not below the 0.05 standard threshold to reject the null hypothesis for each variable. (See Figure 1.1 and Figure 1.2 below).

**Figure 1.1**

*Histograms of PC1 Variables (Grouped by Churn)*

Chart, box and whisker chart

Description automatically generated

**Figure 1.2**

*Code Snippet: Chi-squared Test of PC1 Variables*

Table

Description automatically generated with medium confidence

Table

Description automatically generated with medium confidence

Table

Description automatically generated with medium confidence

## Principal Component 4

The boxplots for both Reliability and Options did not appear to indicate a significant difference between groups. The chi-squared test showed p-values that were well above the 0.05 threshold confirming what the plots showed. (See Figure 2.1 and Figure 2.2 below).

**Figure 2.1**

*Histograms of PC4 Variables (Grouped by Churn)*

*Chart, box and whisker chart

Description automatically generated*

**Figure 2.2**

*Code Snippet: Chi-squared Tests of PC4 Variables*

Graphical user interface, text

Description automatically generated

## Principal Component 6

The boxplots for Children, Age, and Contacts did not show a statistically significant difference in distributions between customers who discontinued service and those that have stayed. The chi-squared tests performed on each confirmed what the histograms implied. (See Figure 3.1 and Figure 3.2 below).

**Figure 3.1**

*Histograms of PC6 Variables (Grouped by Churn)*

*Chart, box and whisker chart

Description automatically generated*

**Figure 3.2**

*Code Snippet: Chi-squared Tests of PC6 Variables*

Graphical user interface, text, application

Description automatically generated

Graphical user interface, text, application

Description automatically generated

## Principal Component 7

The boxplots for both Email and Yearly\_Equip\_Failure did not show a difference between Churn groups, and this was confirmed with chi-squared tests on both variables. (See Figure 4.1 and Figure 4.2 below).

**Figure 4.1**

*Histograms of PC7 Variables (Grouped by Churn)*

*Chart, box and whisker chart

Description automatically generated*

**Figure 4.2**

*Code Snippet: Chi-squared Tests of PC7 Variables*

*Graphical user interface, text, application

Description automatically generated*

# Part III: Conclusions

## Analysis of Variables of Interest

Of the categorical variables analyzed with the chi-squared test, none showed a statistically significant relationship to churn. An interesting observation is that all values of the Customer Satisfaction Survey from the data frame were compared against the Churn variable and no relationship was identified. While the information may be useful, it will not help identify customers that are at risk of churning.

## Limitations on Analysis

The chi-squared test is limited to categorical variables only; therefore, this analysis did not cover all values identified in the Principal Component Analysis. These variables may hold answers to identifying customers who are at risk of churning.

## Additional Data Visualizations

The following data visualizations are included to help address the unanswered question from part 1. The figures 5.1, 5.2, 5.3, and 5.4 help describe the continuous variables Tenure, Bandwidth\_GB\_Year, and MonthlyCharge. The Q-Q plots in Figure 5.1 highlight that Monthly Charge is the only one of the three that follows a somewhat normal distributions (although the tails imply some skewedness). This is confirmed by the histograms in Figure 5.2. The Tenure histogram shows a bimodal distribution which shows a very low number of customers with a tenure between 20 and 40 months. There are a few questions that come to mind after reviewing this graph:

* Did the company offer a new-customer discount that expired after two years?
* Did a new company enter the market that offered discounted or new services that would draw customers away?
* Did something happen that caused customers to lose faith in the company (such as faulty service or a change in leadership to an unknown/unproven candidate)?

The histogram of bandwidth usage per year has a similar bimodal shape with very few customers using 3000-3750 GB per year. Questions from this graph are less obvious because usage is not directly tied to the company and is based more on personal needs. The first peak is for the more casual users and the second is for users who require more bandwidth.

The Figures 5.3 and 5.4 highlight the difference between customers that have churned and those that have not. They show that customers with low tenure and low bandwidth usage tend to leave more than those with high tenure/bandwidth. The scatterplots in Figure 5.4 show this with more detail – customers with a high monthly charge but low tenure *and/or* low bandwidth appear more likely to churn.

The charts in Figure 6.1, 6.2, and 6.3 are included to demonstrate other relationships in the data. Figure 6.2 helps confirm the previous conclusion from the group of Figures 5.1-5.4: the month-to-month contracts appear to have a greater rate of churn than those with one- or two-year contracts. Marital status appears to have no effect on churn based on Figure 6.3.

## Recommended Course of Action

The chi-squared analysis did not reveal a relationship between the categorical variables from the principal components making an answer to the research question unclear. However, the data visualizations of continuous variables *do* seem to imply a relationship between Churn and Tenure, Bandwidth usage, and Monthly Charge. My recommendation would be further analysis of the continuous variables identified in the PCA (paying particular attention to Tenure, Bandwidth, and Monthly Charge) to illuminate any underlying factors that may be contributing to customer churn. It would be logical that a customer who has a high monthly bill but does not use their services (i.e., low bandwidth) and has not been a customer for long may look for service elsewhere.

One conclusion does seem safe to make: after two-year mark (the longest contract offered), the customer is likely to explore other options. The act of looking for a new service as well as the process of transitioning all components of their service to a new company can be an arduous task for some. Offering an incentive to stay with the company may help reduce the number of customers lost.

**Figure 5.1**

*Q-Q Plot of Continuous Variables*

Graphical user interface, chart, line chart

Description automatically generated

**Figure 5.2**

*Histograms of Continuous Variables*

Chart, waterfall chart

Description automatically generated

**Figure 5.3**

*Histograms of Continuous Variables by Churn*

Chart, bar chart, histogram

Description automatically generated

**Figure 5.4**

*Scatterplot of Monthly Charge vs Bandwidth and Tenure, Color-coded by Churn*

Graphical user interface, application

Description automatically generated

Note: The alpha value is for the opacity of each point.

**Figure 6.1**

*Bar Charts of Categorical Variables*

Graphical user interface, chart

Description automatically generated

**Figure 6.2**

*Bar Charts of Categorical Variables by Churn*

Chart, bar chart

Description automatically generated

**Figure 6.3**

*Bar Charts of Contract Type by Marital Status and Area*

Bar chart

Description automatically generated

# Part IV: Supporting Documents

Please find the code file, Panapto recording with my submission, and the references below.

**References**

Jaap & BenBarnes (2012, March 23). How to get summary statistics by Group. Stack Overflow. Retrieved October 28, 2021, from <https://stackoverflow.com/questions/9847054/how-to-get-summary-statistics-by-group>

AltexSoft. (2020, February 27). Customer churn prediction for subscription businesses using Machine Learning: Main Approaches and Models. AltexSoft. Retrieved November 11, 2021, from <https://www.altexsoft.com/blog/business/customer-churn-prediction-for-subscription-businesses-using-machine-learning-main-approaches-and-models/>

Kassambara, A. (2017, September 1). GGPLOT2 - easy way to mix multiple graphs on the same page. STHDA. Retrieved October 28, 2021, from <http://www.sthda.com/english/articles/24-ggpubr-publication-ready-plots/81-ggplot2-easy-way-to-mix-multiple-graphs-on-the-same-page/>

One-way ANOVA test in R. STHDA. (n.d.). Retrieved October 28, 2021, from <http://www.sthda.com/english/wiki/wiki.php?title=one-way-anova-test-in-r>

Stacked bar chart in GGPLOT2. R CHARTS | A collection of charts and graphs made with the R programming language. (2021, March 27). Retrieved November 12, 2021, from <https://r-charts.com/part-whole/stacked-bar-chart-ggplot2/>

Unpaired two-samples Wilcoxon test in R. STHDA. (n.d.). Retrieved October 27, 2021, from <http://www.sthda.com/english/wiki/unpaired-two-samples-wilcoxon-test-in-r>